高能物理领域科学数据复用特征及影响因素研究

胡威 [1,2]　杨宁 [1]

[1] 中国科学院成都文献情报中心　成都　610041

[2] 中国人民大学统计学院　北京　100872

摘要：[目的/意义]以 Data Citation Index（DCI）数据库高能物理领域科学数据为研究对象，探究高能物理领域科学数据的复用特征及影响因素，为推动我国数据共享和引用规范性、提升数据价值和影响力提供参考与借鉴。**[方法/过程]**利用 DCI 数据库的数据基本信息和引用信息，采用统计回归方法，通过科学数据属性特征、科学数据复用特征、科学数据属性特征与复用特征相关性 3 个维度开展高能物理领域科学数据复用特征及影响因素的分析。**[结果/结论]**研究结果表明，高能物理领域科学数据共享数量逐年递增，但数据字段缺失比例较高，数据复用受数据等级、出版模式和学科类别的影响较大，导致被引频次分布极不均匀，高等级科学数据更易获得高复用次数，科学数据共享和引用规范有待进一步加强。最后，本文据此提出高能物理科学数据复用的优化提升路径。

关键词：科学数据；高能物理；数据复用；影响因素；统计回归

分类号：G350

Research on Scientific Data Reuse Features and Influencing Factors in the Field of High Energy Physics

Hu Wei[1,2], Yang Ning[1]

[1]*National Science Library(Chengdu), Chinese Academy of Sciencess, Chengdu* 610041;

[2] *School of Statistics, Renmin University of China, Beijing* 100872

Abstract: [Purpose/significance] By utilizing the Data Citation Index (DCI) database, this article explores the reuse features and influencing factors of scientific data in the field of high-energy physics. These findings serve as a point of reference and support, facilitating the promotion of data sharing and citation standardization in China. Moreover, these contribute to the augmentation of both value and influence of scientific data.**[Method/process]** This article adopt statistical regression methods to analyze the basic and citation features of the DCI database. For the reuse features and influencing factors, the analysis includes three dimensions: scientific data attribute features, reuse features, and correlation between attribute and reuse features. **[Result/conclusion]** The research findings reveal that the publication volume of scientific data in the field of high-energy physics is exhibiting an increasing trend. However, the proportion of missing data fields is relatively high. The reuse of high-energy physics scientific data is significantly influenced by publication modes and disciplinary categories. These result in the extremely uneven distribution of citation frequency. High-level scientific data are more likely to be reused. Moreover, the standardization of scientific data sharing and citation needs further enhancement. Finally, we propose an optimization and improvement path for high-energy physics science data reuse based on this findings.

Key words: Scientific data; high-energy physics; data reuse; influencing factors; statistical regression model

*基金项目：本文系中国科学院文献情报能力建设专项项目“支撑院党组决策的战略情报感知平台建设与应用”（编号：292021000479）和成都市软科学研究项目“成都市深化科技体制改革的政策研究——以科学数据中心的建设为路径”（编号：2023-RK00-00061-ZF）的研究成果。

1 引言

科学数据复用（Data Reuse）也被称为“数据重用”“数据再利用”，指的是科研人员为了重现研究结果或新的研究目的而对数据进行二次使用的行为[1]。随着数据密集型科研范式的兴起和发展，科研活动过程中所产生的科学数据已经成为一项重要的科研成果产出，对其进行共享和复用也变得日益普遍[2]。数据复用可以节约科研人员的研究成本、缩短科研周期、增加科研产出，同时还可以从侧面扩大科研仪器共享范围、减小学术造假行为发生的可能性，数据复用给科学研究带来的价值和贡献已经成为学界共识[3,4]。

科学数据复用实践在自然科学领域的学科中开展较早，如生命科学[5-7]、工程科学[8]、地球科学[9]、信息科学[10]等数据驱动型学科。随着科学数据参与科研活动日渐广泛，社会科学领域如图书情报[11]、管理学[12]、经济学[13]等学科也开始围绕科学数据复用开展研究和实践。近年来，相关学者主要从两个视角对科学数据的复用展开研究。①科学数据复用管理者的视角。在管理者视角下，科学数据基础设施建设[14]、数据复用的政策环境[15]、数据复用的标准制定[16]、数据复用的模型框架[17]等都成为研究热点。20 世纪 90 年代开始，在数据开放共享与重用需求的推动下，科学数据复用政策不断出台[18,19]，数据基础设施建设也持续推进[20,21]，数据标准与元数据方案陆续制定[22-24]，这些管理政策和标准增强了科研人员在数据复用过程中尊重数据生产者权益的意识，从数据复用管理者角度规范并促进了数据复用活动的开展。②科学数据复用参与者的视角。图书情报机构[25]、科学数据中心[26]、出版商[27]及科研人员等作为科学数据复用活动流程的参与者，目前均已产生大量研究。尤其是围绕科研人员数据复用行为的研究，包括在数据复用过程中的数据获取、处理、使用与评价[28]，以及对数据复用的态度[29]、意愿[30,31]、影响因素[32,33]等研究最为热门。权衡科学数据复用全生命周期中各参与者的权益问题，是科学数据复用活动长期开展的根本保障[34]。

高能物理又名粒子物理，主要研究比原子核更微观层级的物质结构与性质，以及在极高能量条件下物质相互转化的规律，是典型的数据密集型学科。由于高能物理科学数据的多源性和复杂性，当前研究主要围绕数据平台建设[35,36]、数据服务[37]、文献与数据关联[38]及开放共享机制[39]开展，对于高能物理科学数据复用特征及影响因素的探究还鲜有开展。因此，本研究以 Data Citation Index（DCI）数据库高能物理领域科学数据为研究对象，通过统计回归方法，对高能物理领域科学数据的复用特征及影响因素进行深入分析，为推动我国高能物理领域科学数据共享和引用规范性、提升我国科学数据的价值和影响力提供参考。

2 数据来源与研究方法

高能物理科学数据的收录范围十分广泛，包括科学实验的过程记录、结果记录、模拟实验数据、描述数据和指示数据存储位置等信息的元数据等，这些数据都与最终的研究成果密切相关[40]。按照数据产生来源的不同，高能物理领域的科学数据可以分为实验数据和模拟数

据。其中，实验数据是在实验装置中采集及经过后续加工处理得到的数据，模拟数据是基于物理模型通过计算机模拟计算得到的数据。为了促进高能物理科学数据的长期保存和开放共享，欧洲核子研究中心 (CERN) 、美国布鲁克海文国家实验室 (BNL)及中国科学院高能物理研究所 (IHEP)等数十家国际高能物理研究机构于 2008 年共同成立了国际高能物理数据长期保存组织（DPHEP）[41]，该组织在 2012 年发布了高能物理数据长期保存蓝图，定义了高能物理数据共享的四个等级，其含义如表 1 所示。

表 1 高能物理科学数据共享的四个等级及其含义

数据等级	含义	备注
Level-1	公开文档或论文数据	最终的结果数据，原始数据积分、求和后的结果，无法通过逆向操作得到原始数据。
Level-2	用于教育或科普的简单格式数据	对原始数据进行处理后的简单格式数据，无需特定软件即可进行数据分析，但不能支持完整的物理学分析。
Level-3	支持完整科学分析的重建数据	保存能够支持完整物理学分析的重建数据，数据保存和软件维护的成本高于 Level-2。
Level-4	所有原始数据及相关条件数据和软件	能够支持重现和生成新的真实数据和/或模拟数据，可以获得所有的潜在数据。在数据保存方面的要求最为严苛。

2.1 数据来源

本文的研究数据收集自 Web of Science（WOS）的 DCI 数据库，该数据库是 WOS 的子库之一，由汤森路透公司于 2012 年推出，经过不断地迭代更新后，目前可以支持一站式地进行科学数据资源的检索、获取和研究，并且建立了科学数据及其对应研究论文的关联。DCI 数据库较好地记录了科学数据的出版和复用情况。以科学数据的被引频次（Z9）字段为例，DCI 数据库统计了包括来自 WOS 核心合集在内的 7 种数据库中收集到的数据被论文引用的情况。

本研究以 Physics, Particles & Fields （粒子物理）、Astronomy & Astrophysics（天文与天体物理学）作为学科类别进行检索，最终下载并整理后得到数据 172684 条。从 DCI 数据库下载并提取得到的字段包括：科学数据的可访问链接、数字对象标识符（DOI）、标题、关键词、学科、出版年份、出版模式、出版机构、数据类型、作者、作者地址、团体作者、所属机构、国家/地区、语种、被引频次（包括 WOS 核心合集、来自 7 种数据库、来自论文的引用、来自相同学科的引用等细分）、引用的参考文献数量、学科类别等。

2.2 研究方法

本文总体研究思路是基于 DCI 数据库的高能物理领域科学数据，利用统计分析方法从高能物理领域科学数据的属性特征和复用特征两个维度，对高能物理领域科学数据的总体情

况、出版模式、从属情况、质量情况、复用情况和高被引数据等进行研究分析。其次，利用回归分析方法对高能物理科学数据属性特征与复用特征的相关性进行分析，主要包括数据属性特征与复用影响分析、数据出版与等级对复用影响分析两个维度。最后，利用分析结果辨析当前高能物理科学数据复用存在的问题和难点，并提出相应的对策建议。整体研究思路如图 1 所示。

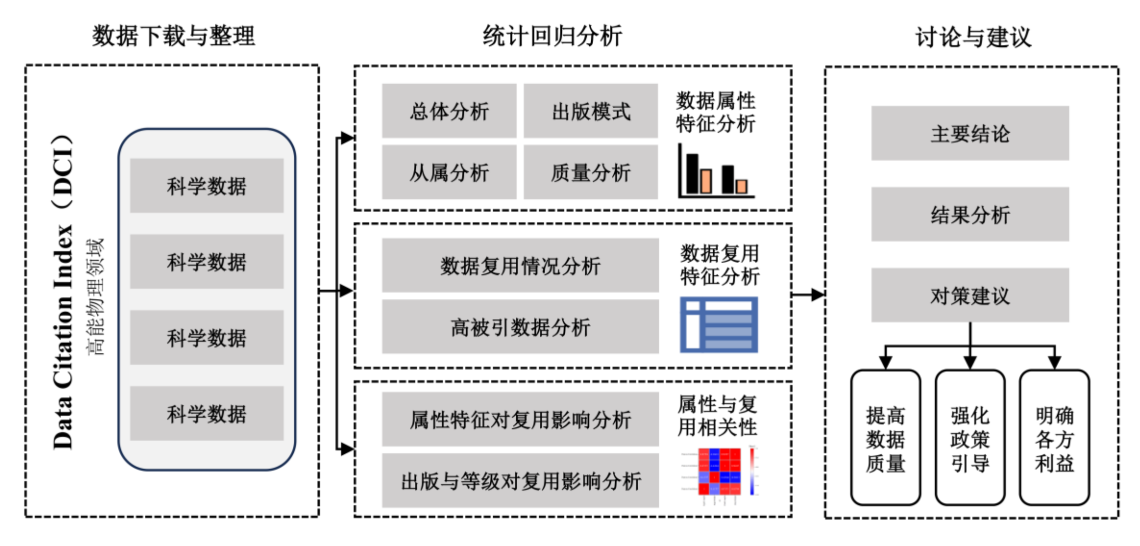


图 1　本文整体研究思路及过程

在模型选取方面，本研究以复用次数作为因变量Y，因其属于计数型数据即取值为非负的整数数据，并且高能物理科学数据复用次数为 0 的比例较高，导致复用次数标准差远大于均值，最终结合不同模型的拟合优度，本文选择 0 膨胀泊松回归模型进行相关性分析[42]。回归模型的自变量X为高能物理科学数据的属性特征，包括出版模式、从属情况、质量情况、发布时间以及前文所提及的高能物理科学数据等级。

0 膨胀泊松回归模型通过加入伯努利分布以拟合解释高 0 值比例的现象，该模型可以表示为一个伯努利分布和一个泊松分布的混合模型，记$Y \stackrel{\text{def}}{=} ZY'$，其中$Z$服从参数为 $1-\pi$的伯努利分布，Y'服从参数为λ的泊松分布，Z和Y'这两个分布都会生成 0 值。因此，因变量Y的概率分布表示为：

$$P(Y=y)=\begin{cases}\pi+(1-\pi)\exp(-\lambda), y=0\\(1-\pi)\dfrac{\lambda^y\exp(-\lambda)}{y!}, y>0\end{cases}$$

其中，参数π和λ写为自变量X的回归模型，即$\log(\lambda)=X\beta$,$\text{logit}(\pi)=X\gamma$,$X=\left(X_1,\ldots X_p\right)\in R^p$表示$p$个自变量，$\beta^{\top}\in R^p$和$\gamma^{\top}\in R^p$表示对应的回归系数，$\text{logit}(\pi)=\log(\pi/(1-\pi))$。

在结果解读方面，各个自变量回归系数的绝对值大小表示了对于复用次数影响的相对重要程度，绝对值越大则表示该特征的影响越大。回归系数β（泊松部分的回归系数）、γ（伯努利部分的回归系数）分别表示在保持其他因素不变的条件下λ的对数、π的对数几率的变化幅度，再经过共线性检查和 BIC 变量选择，可以得到最终回归结果。对自变量X_j可以根据其

系数估计值及显著性，解读如下：在保持其他因素不变的条件下，(1) 若$\gamma_j < 0$或$\beta_j > 0$，则表示自变量X_j的增加或分类变量取值为某类别时，会使得数据复用次数提高；(2) 若$\gamma_j > 0$或$\beta_j < 0$，则表示自变量X_j的增加会使得数据复用次数降低；(3) 若γ_j与β_j均显著不为 0，且符号同正或同负，则需将系数估计值代入Y的概率分布判断自变量对数据复用次数的影响方向；(4)泊松部分β系数表示对于复用次数大于 0 的科学数据（即$Y_i > 0$），自变量X_j的增加对于科学数据能够取得更多复用的可能性。

2.3 变量选取与定义

为描述科学数据的基本属性特征并研究科学数据复用特征及影响因素，本文以被引频次（即数据的复用次数）作为回归模型的因变量 Y，并通过数据预处理及观察测试，得到可能对数据复用次数影响较大的出版模式、从属情况、质量情况、发布时间及科学数据等级作为自变量X。本文选取的变量及详细定义如表 2 所示。

表 2　本文选取的变量及说明

变量类别	变量名称	变量类型	变量定义与取值范围
因变量	被引频次（复用次数）	计数型变量	DCI 数据库中的“被引频次合计”字段，取值范围从 0 到 571
出版模式及等级	出版模式	多分类型变量	Data set 等 4 种出版模式，基准组为 Data set
	数据等级	多分类型变量	若科学数据的标题中出现 Table 等字样则记为“表格型科学数据”；若标题中出现 Figure 等字样则记为“图片型科学数据”；若出版模式为 Data set 而且非图片、表格型科学数据，则记为“简单格式科学数据”，基准组为“图片型科学数据”
从属情况	所属机构	多分类型变量	HEPData 等 24 种出版机构，基准组为其他
	学科类别	多分类型变量	共有 Physics, Particles & Fields 等 14 个学科，其中部分科学数据对应多个学科，将出现频次小于 10 的学科合并为“其他”类，并作为基准组
质量情况	缺失的字段数目	连续型变量	单位为个，DCI 数据库中缺失的字段个数,取值范围从 2 到 6
其他	发布时间	连续型变量	单位为年，即该科学数据发布年份，取值范围从 1900 到 2022

3　研究结果分析

3.1 高能物理领域科学数据属性特征分析

3.1.1 数据总体情况分析

首先对 DCI 数据库中高能物理科学数据的逐年收录情况进行统计分析，得到 1980-2022 这 40 余年间高能物理科学数据逐年收录数量，数据总体情况如图 2 所示。

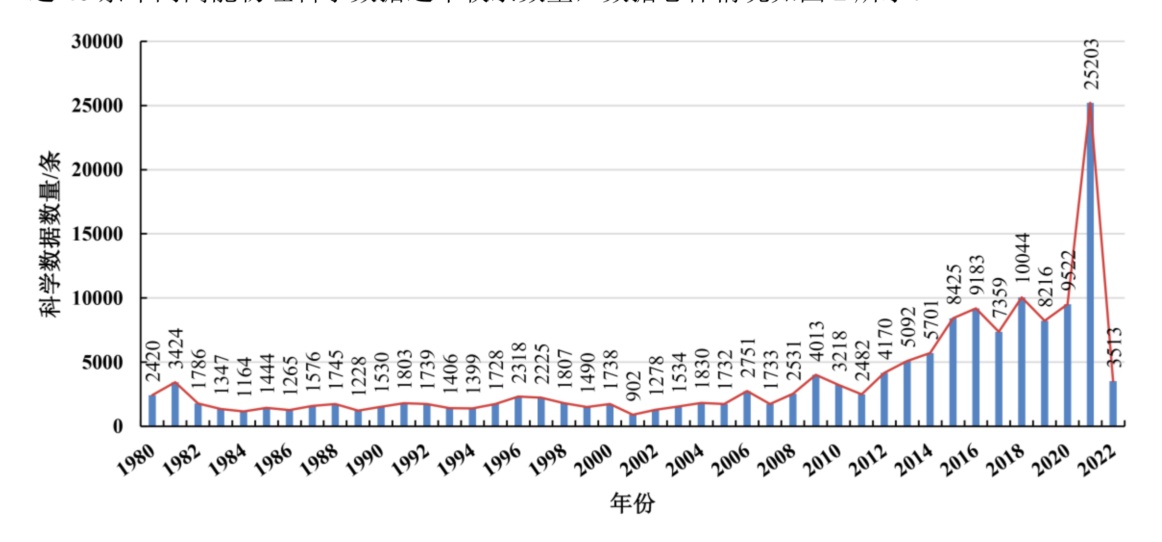


图 2　DCI 高能物理科学数据逐年收录数量分布

总体来看，DCI 收录的高能物理科学数据在 1980 年只有 2420 条，而到 2021 年已经达到 25203 条，增长了 941.5%，年均增长 22.4%，增长趋势十分明显。从图中也可以清晰地发现，高能物理领域科学数据的收录数量呈现波动式上升的态势，并在 2006 年后开始急剧增长，这与科研范式的转变时间基本吻合，说明科学数据作为科研过程中的一项重要成果产出，其共享和复用正变得日益普遍。同时也说明随着高能物理领域的快速发展和资源的持续投入，其研究成果的数量呈现持续上升的趋势，研究热点不断涌现。

3.1.2 数据出版模式分析

DCI 数据库对数据进行了多层级的信息标引，并通过出版模式（DT）字段标注了数据所属的层级和模式。通过数据分析发现，高能物理领域科学数据共分为四种出版模式，分别为数据集（Data Set）、软件（Software）、数据研究（Data study）和数据仓储（Repository）。第一种出版模式为 Data Set 的科学数据包括简单格式数据、数据的基本描述信息、处理后得到的表格数据或图片数据等，其数量占比最高；第二种出版模式为 Software 的科学数据通常表现为软件程序包的形式，同时还附带有示例数据和使用说明文档，可以帮助研究者更方便地用其分析和处理其他数据；第三种出版模式为 Data study 的科学数据通常是将数据集与相关的科学研究描述文件、过程记录相结合，以便更深入地了解数据的背景、来源和含义，强化了科学数据与科学研究之间的关系和关联性；第四种出版模式为 Repository 的科学数据指的是用于储存和管理数据集和软件程序包的规范化数据仓储设施，更关注于数据的可靠查找、访问和管理服务。四种出版模式的层级关联关系及数据数量如图 3 所示。

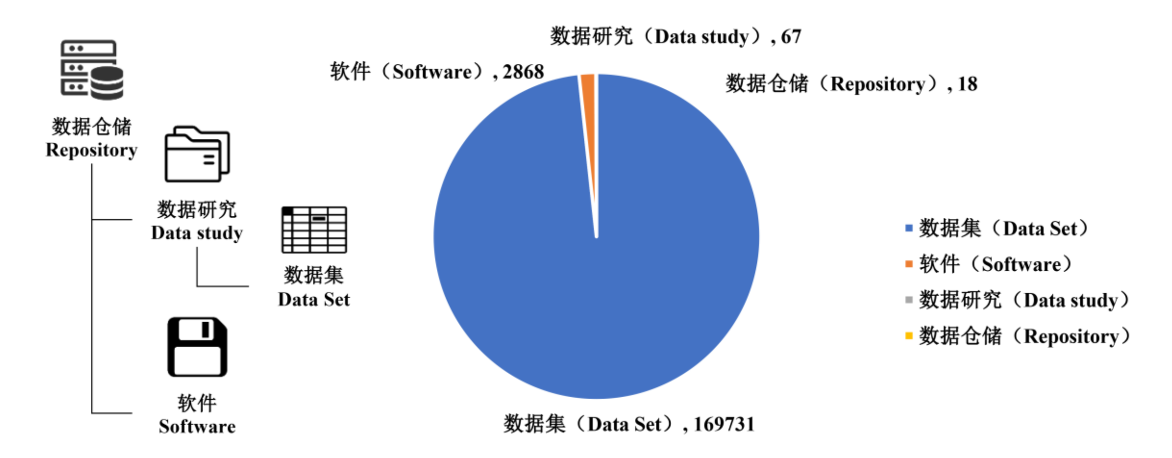


图 3　DCI 高能物理科学数据出版模式层级关系及数据数量

由图可知，在 DCI 高能物理科学数据中出版模式标注为数据集（Data Set）的数量占绝大多数，比例高达 98.3%。其次是软件（Software），共 2868 条数据，占比约为 1.7%。这是由于数据集和软件是高能物理科学数据出版的最终形式，其作为独立个体发布的数量较多，数据共享发布过程也较为简单。而收录了高能物理研究过程信息的数据研究（Data study），需要将科研过程的临时数据、描述文件和日志记录等信息进行整理、整合和发布，共享发布流程较为复杂，科研人员更习惯将其中数据进行独立发布，而非整合后发布。而数据仓储（Repository）与高能物理领域数据仓储的建设数量相关，这种出版模式的数据量相较于其他三种模式较少。

3. 1. 3 数据从属情况分析

对 DCI 数据库高能物理科学数据的作者、团体作者、所属机构和学科类别进行分别统计分析，了解数据的从属情况。通过计算后发现，共有超过 200 位高能物理领域研究者参与了超过 10000 条科学数据的研究和共享工作，DCI 数据库中高能物理科学数据中最多作者的数量达到 2840 人。统计计算结果也表明，大部分科学数据是由少部分研究者贡献的，符合普赖斯定律。

表 3　DCI 高能物理科学数据高频团体作者、所属机构与学科类别

团体作者	数量	所属机构	数量	学科类别	数量
CMS	17070	ILL	3080	粒子物理	149834
ATLAS	16749	JINR	2281	天文与天体物理学	23747
STAR	7712	CERN	1846	科技其他主题	5612
ALICE	3462	SERPUKHOV IHEP	1588	光学	893
ZEUS	2455	ANL	1280	心理学	108

DCI 高能物理科学数据高频团体作者及所属机构如表 3 所示。从团体作者角度来看，科学数据的团体作者十分署名很常见，说明高能物理领域的科学研究团队合作程度较高。此外，在团体作者署名中，每个作者都对研究结果负有责任和义务，这有助于确保研究结果的可信度和可重复性。发布数据最多的团体为 Compact Muon Solenoid（CMS）Collaboration，该机构是全球最大的科学合作组织之一，汇集了来自 50 多个国家的约 240 个研究所和大学的粒子物理学家、工程师、计算机科学家、技术人员和学生，有超过 3000 名研究人员参与其中。从所属机构角度来看，发表数据最多的机构为劳厄-朗之万研究所（ILL），该机构位于法国格勒诺布尔，由法国、德国和英国与其他 11 个欧洲国家合作资助和管理，其他发布数据较多的机构还包括俄罗斯杜布纳联合核研究所(JINR)、欧洲核子研究组织（CERN）等。

DCI 高能物理科学数据所属学科类别情况来看，数据属于多学科交叉的情况并不十分常见，绝大多数数据都标注了明确的学科类别。其中，粒子物理数据为 149834 条，占比高达 86.8%，排名第一。数量较多的学科还包括天文与天体物理学、科技其他主题及光学等，其他还出现了心理学、生命科学等交叉学科。

3.1.4 数据质量情况分析

对 DCI 高能物理科学数据的缺失情况进行统计，进而分析数据的质量情况，结果如表所示。从中可以发现三个主要问题：首先，部分重要字段缺失比例较高，如关键词（DE）字段，其缺失比例达 91.4%，关键词对于数据检索和发现有着极其重要的意义，完善的关键词有助于提高科学数据被发现和复用的概率。其次，对于作者机构、地址等重要信息，普遍存在未统一和消歧的情况，数据填写较为随意，导致统计分析时产生较大误差。最后，字段含义不清晰，以数据类型（DY）字段为例，其缺失比例达 80.0%，且含义不明确，其中既有“Scattering Data”表示数据类型是散点型，也有“Astronomical Data” 表示数据属于天文学领域，还有“Monte Carlo Simulation” 表示该数据是蒙特卡洛随机模拟数据。这说明该字段的定义不够清楚和明确，导致作者填写的信息类型含义不一致或漏填。

表 4 DCI 高能物理科学数据部分字段含义及缺失比例

字段	含义	缺失比例	字段	含义	缺失比例
UR	该科学数据的网站链接	97.7%	PY	出版年份	0
DE	科学数据关键词	91.4%	SO	出版机构	0
DY	数据类型	80.0%	SU	学科	0
C1	作者机构及地址	78.2%	TI	标题	0
CA	团体作者	54.9%	U1	使用次数(近 180 天)	0
MI	数据表格的标注标签	36.6%	U2	使用次数(2013 年至今)	0
DI	数字对象标识符(DOI)	2.3%	UT	入藏号	0
DT	出版模式	0	WC	WOS 学科类别	0

LA	语种	0	Z9	被引频次合计	0
NR	引用的参考文献数量	0	PT	出版物类型	0

3.2 高能物理领域科学数据复用特征分析

3.2.1 数据复用情况分析

通过科学数据的被引频次（Z9）字段对科学数据的复用情况进行研究，将被引频次作为科学数据复用次数进行计算分析。经过研究发现，有 86.2%的科学数据的复用次数为 0，有 13.7%的科学数据复用次数为 1，而复用次数大于 1 的科学数据仅占 0.1%，DCI 高能物理科学数据复用次数分布情况如图 4 所示。

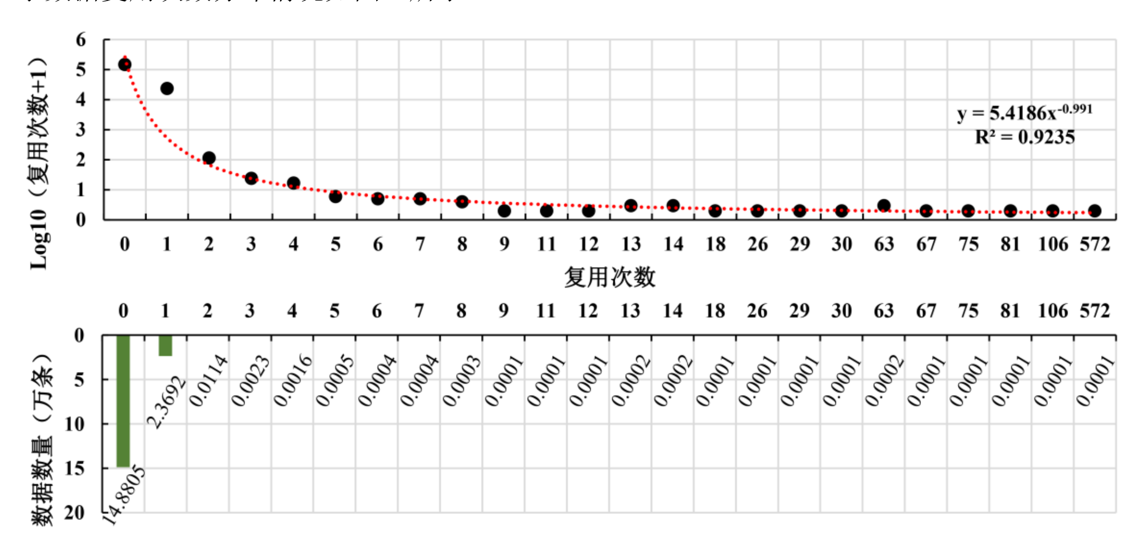


图 4　DCI 高能物理科学数据复用情况

由图 4 可见，高能物理科学数据的复用次数分布极不均匀，绝大多数的科学数据的复用次数很低，而极少数科学数据的复用次数很高，这与期刊论文的被引频次分布规律是一致的。如图 4 也可知，经对数处理后的复用次数，其幂率分布的拟合优度R^2=0.9235，近似服从幂律分布。从整体情况来看，高能物理领域科学数据的整体复用比例较低，平均复用次数为 0.15 次，这表明高能物理领域科学数据的复用仍有较大的提升空间，相比于生物医学等领域科学数据被复用的次数偏低[43]。

3.2.2 高被引数据分析

对高能物理科学数据的高被引情况进行分析发现，在所收集的 172684 条数据中，仅有 10 条数据的复用次数高于 20 次。DCI 高能物理领域排名前 5 的高被引科学数据的具体情况如表 5 所示，其中“复用次数”指能够从 Web of Science 核心合集、Arabic Citation Index、BIOSIS Citation Index、中国科学引文数据库、Data Citation Index、Russian Science Citation Index、SciELO Citation Index 这 7 个数据库中收集得到的数据被引频次之和，除论文引用以外，还包括来自专利、报告、软件等出版物的引用。“来自论文引用的频次”指来自论文而

非其他出版物的引用频次，排除了其他渠道引用数据的影响。“来自本领域的引用频次”指来自于高能物理领域出版物的引用次数，既包括来自高能物理领域论文的引用次数，也包括来自高能物理领域的专利、报告、软件等其他出版物的引用次数。

表 5　DCI 高能物理高被引科学数据情况

科学数据标题	年份	出版模式	学科类别	复用次数	来自论文的引用频次	来自本领域的引用频次
ATNF Pulsar Database	2003	数据仓储	天文与天体物理学	571	421	422
Sloan Digital Sky Survey SkyServer SDSS Data Release 1 (DR1)	2003	数据研究	天文与天体物理学	106	90	90
pynbody: NBody/SPH analysis for python	2013	软件	天文与天体物理学	30	30	30
A Study of Bhabha Scattering at PETRA Energies	1988	数据集	粒子物理	29	29	29
Sloan Digital Sky Survey	2000	数据仓储	天文与天体物理学	26	15	6

由上表可见，DCI 高能物理科学数据复用次数最高的是“ATNF Pulsar Database”，该数据是出版类型为数据仓储的科学数据，该数据仓储收集了所有已知旋转动力脉冲星和磁星的基本参数，由澳大利亚国家科学机构 CSIRO 的空间和天文学业务部门运营和管理，保持持续的维护更新，并提供相关的代码工具供世界各地机构的研究人员访问使用。排名第五的“Sloan Digital Sky Survey”同样为数据仓储类型的科学数据，该数据为美国的斯隆数字巡天项目数据库，数据来自位于美国新墨西哥州阿帕奇山顶天文台的 2.5 米口径望远镜，该项目也被称为全世界最成功的巡天计划之一。其他 3 个数据集分别为出版类型为数据研究、软件和数据集的科学数据，包括同样来自于斯隆数字巡天项目发布的数据研究类型的数据、天体物理模拟的分析框架软件（pynbody）以及 TASSO 合作组织对 Bhabha 散射多年的研究数据集。

通过高被引科学数据的情况分析可以看出，持续性的科学数据运营管理和维护更新是数据被高复用的关键必要条件。此外，天文与天体物理学的科学数据数量虽然只有粒子物理科学数据的 1.6%，但在高被引数据中确占比和排名均明显高于粒子物理，得到了较多的引用次数。

3.3 科学数据属性特征与复用特征相关性分析

3.3.1 数据属性特征对复用影响分析

通过 0 膨胀泊松回归模型，本文对高能物理科学数据的属性特征与复用特征之间的相关关系进行总体情况分析，据此探究复用的显著影响因素，并度量科学数据等级、出版模式、从属情况、质量情况对于复用情况的影响程度。经过计算，本文所得到的 0 膨胀泊松回归模型结果如**错误!未找到引用源。**所示，其中省略了截距项的估计结果，系数上标“***”表示在 0.001 水平下显著不为 0，“###”表示经 BIC 变量选择后该自变量未被纳入最终模型，或因在设定的 0.001 水平下不显著而省略。

表 3　0 膨胀泊松回归模型的系数估计结果

自变量类别	自变量名称	系数估计结果	
		伯努利部分γ	泊松部分β
出版模式及等级	数据仓储（Level-3 或 4）	###	4.979***
	软件（Level-3 或 4）	###	2.882***
	数据研究（Level-3 或 4）	-6.997***	2.592***
	简单格式数据（Level-2）	###	###
	图片型数据（Level-1）	###	###
	表格型数据（Level-1）	###	###
从属情况：学科类别	天文与天体物理学	###	4.420***
	粒子物理	###	###
从属情况：所属机构	Astrophysics Source Code Library	###	-2.751***
	Centre De Donnees Astronomiques De Strasbourg	###	-8.672***
	CERN Open Data Portal	###	3.297***
	HEPData	###	3.776***
	Institut Laue-Langevin	###	1.216***
	Planetary Data System	-10.986***	-1.800***
质量情况	缺失的字段数目	###	###
时间效应	发布时间	9.164***	0.314***

对**错误!未找到引用源。**中模型系数估计结果进行解读，可得到科学数据属性特征对于复用次数的影响。其中系数β对应于复用次数大于 0 时，自变量对于复用次数的影响。泊松部分β_j系数的含义可解释为：在保持其他因素不变的条件下，如果β_j值越大，那么自变量X_j

对应的类别，就更有可能使得科学数据获得更多复用次数并成为高被引数据。最终，根据以上系数估计结果，可得出以下结论：

（1）对于出版模式及科学数据等级的层面，相比较于数据等级为 Level-1 的图片型数据和表格型数据，高等级的科学数据因其规范化的仓储设施、更丰富的数据背景信息以及相关软件的介绍，从而更易获得更多的复用次数。例如，对于复用次数大于 0 的科学数据，若出版模式为数据仓储，则在保持其他因素不变的条件下，该科学数据的平均复用次数比图片型科学数据高 4.979 次。

（2）对于从属情况中的学科层面，学科类别聚焦于“天文与天体物理学”的科学数据的复用可能性比粒子物理、多学科类别、或不确定学科类别归属的科学数据更高。对于从属情况中的所属机构层面，所属机构与出版模式和学科类别自变量之间存在相关性，不同机构会在出版模式和学科类别的选择上有所偏好，表现出一定规律性。

（3）对于数据质量的层面，在保持其他因素不变的条件下，字段缺失数量对于复用的影响并不显著。字段缺失数量与复用次数之间的皮尔逊相关系数<0.1。根据进一步的描述分析发现，这是因为不同类型、学科的科学数据之间存在着字段缺失数量的差异，而不是字段缺失数量本身对复用有显著影响。

从科学数据管理层面考虑，科学数据中心、出版机构本身对于提高高能物理科学数据的复用比例及次数均存在着迫切需求。为研究科学数据复用的优化提升路径，本文对各自变量对于数据复用次数的平均正向影响程度进行刻画，结果如图 2 所示。出版模式及科学数据等级这一类自变量对于复用次数的平均影响强度为 3.48，在四类自变量中能够占据 30%的影响力。而发表时间、学科类别与所属机构均由科学数据作者自主选择，相对来说较为客观，非数据管理者的可控因素。因此，提高科学数据等级将会是优化和提升科学数据复用的有效路径。科学数据管理者可主动利用计算资源通过构造数据论文关联关系等方式，提高科学数据等级，以增加其被复用的可能性。

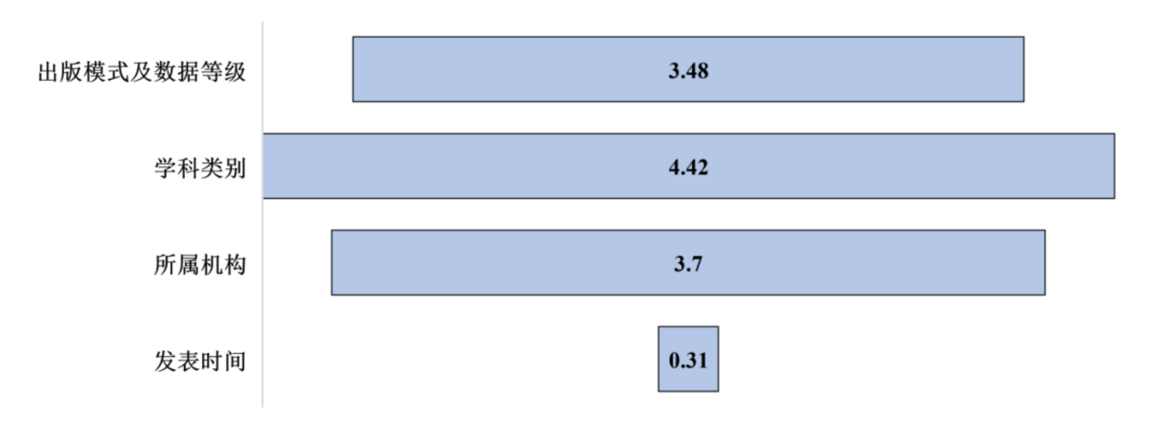


图 2　不同自变量对数据复用次数的平均正向影响程度

3.3.2 数据出版与等级对复用影响分析

通过 0 膨胀泊松回归模型，得到了不同属性特征对于复用次数影响的总体情况，探究了能够提高平均复用次数的影响因素。为进一步细致比较出版模式及数据等级对于数据复用次数的影响，本文首先对 4 种出版模式的科学数据复用情况进行研究。由于 0 取对数无意义，故对复用次数+0.01 后再取对数，得到的复用次数分布如图 3（a）所示，将纵坐标向上平移 2 个单位使得最小取值为 0。由图中可以看出，出版模式为数据仓储（Repository）的科学数据的平均复用次数最高，其次是数据研究（Data study），远高于软件（Software）和数据集（Data Set）。其中各箱体的叉号代表平均复用次数，箱体内部的黑色横线代表中位数，各出版模式的科学数据复用次数表现为极不均匀的右偏分布，因此平均复用次数大于复用次数中位数。另外，由于软件和数据集这两种出版模式的科学数据，其复用次数的中位数均为 0，所以中位数黑色横线与箱体的下横线重合。

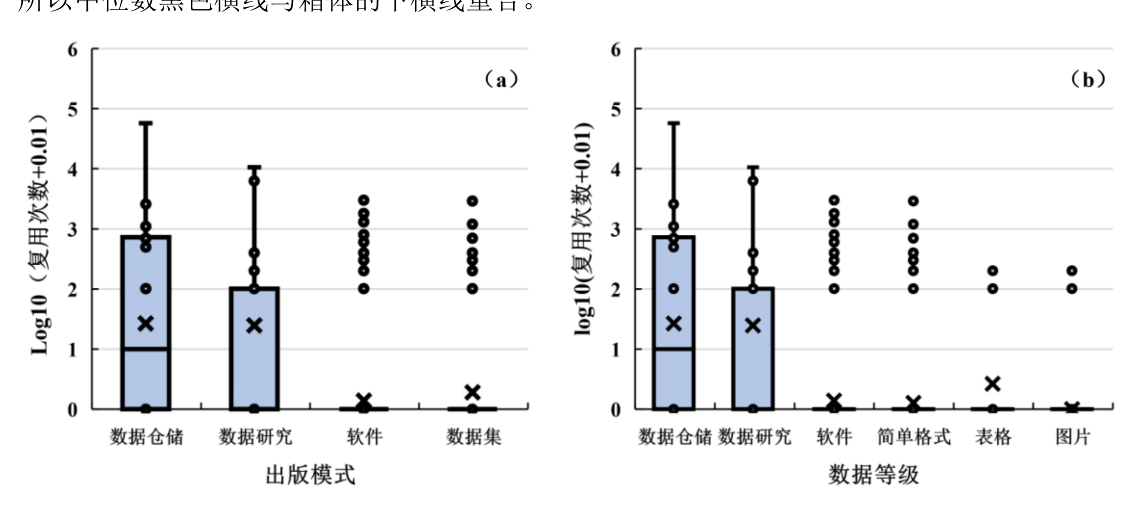


图 3　不同出版模式和等级的高能物理科学数据复用情况

结合科学数据等级、数据格式和出版模式，科学数据可以被进一步细分为 6 种类型，包括 Level-1 的表格和图片型数据集、Level-2 的简单格式数据集、Level-3 或 Level-4 的数据研究、软件和数据仓储数据，得到结果如图 3（b）和表 4 所示。由图表可以得出两方面的结论：（1）从平均复用次数的层面，属于 Level-1 等级的表格数据平均复用次数要高于同为 Level-1 等级的图片型数据，甚至高于等级更高的简单格式数据和软件。（2）从高被引的层面，软件和简单格式数据中高被引数据的个数及复用次数均多于表格和图片数据，即科学数据的等级越高，越容易出现高被引数据。

表 4　不同等级及出版模式的复用情况描述性分析

科学数据等级	出版模式	样本量	均值	标准差	50%分位数	99%分位数
Level-1	数据集：表格型数据	95527	0.211	0.408	0	1
Level-1	数据集：图片型数据	7998	0.002	0.046	0	0
Level-2	数据集：简单格式数据	66215	0.056	0.277	0	1

Level-3 或 4	数据研究	67	7.540	22.071	1	89.500
Level-3 或 4	软件	2859	0.135	0.946	0	2.420
Level-3 或 4	数据仓储	18	35.100	134.149	0.500	479

为了检验表 4 中，六种科学数据等级及出版模式的复用次数之间的差异是否显著，本文将其两两为一组进行被引频次的独立样本均值检验。原假设表示两种科学数据的复用次数无显著差异。由于复用次数呈现极不均匀的非正态分布，因此采用 Wilcoxon 秩和检验方法，假设检验的p值结果见表 5。根据表 5 的结果，可认为在 0.01 的水平下：（1）整体来看，各科学数据等级之间的复用次数的差异显著，结合统计回归的结果，预期提高科学数据等级能够显著提高复用次数；（2）对于科学数据等级为 Level-3 或 4 的数据研究和数据仓储，这两种出版模式的科学数据的复用次数差异不显著。

表 5　复用次数差异的显著性检验结果

出版模式	表格型数据	图片型数据	简单格式数据	软件	数据研究	数据仓储
表格型数据	--	<0.001	<0.001	<0.001	0.008	<0.001
图片型数据	--	--	<0.001	<0.001	<0.001	<0.001
简单格式数据	--	--	--	0.002	<0.001	<0.001
软件	--	--	--	--	<0.001	<0.001
数据研究	--	--	--	--	--	0.887
数据仓储	--	--	--	--	--	--

4　讨论与建议

通过本文分析结果来看，高能物理领域科学数据具备多源、异构、复杂等特点，导致数据共享和复用仍然存在着很大提升空间。首先，高能物理科研人员更习惯于在领域数据仓储或数据平台中共享和复用数据，如 INSPIER、HEPDATA、Zenodo 等，而 DCI 等多学科科学数据共享和索引平台并未得到广泛应用，国内只有清华大学图书馆等少量机构开通了该数据库的访问权限，导致无法建立起统一的引用标准和规范，影响了数据的传播和复用。其次，高能物理科学数据通常容量较大，如 Level-2 以上级别的数据可能达到 PB 级以上，这给数据的保存和共享带来了极大困难，需要投入较多成本用于存储和网络设备的建设。最后，高能物理科学数据质量层次不齐，部分重要字段如作者、关键词等存在较多缺失，这给科研人员造成较大困扰，或因无法判断数据的完整性导致数据被复用的可能性进一步减小。

4.1 提升科学数据等级，实现高能物理科学数据价值，增加可复用性

高能物理科学数据的分级分类出版模式给复用数据带来了较大便利，科研人员和数据仓储可主动发挥管理效能，利用现有技术手段，提高高能物理科学数据的出版模式和数据等级。根据实证研究结果，提高科学数据能级是能够显著加速提高科学数据复用次数的可行手段。首先，对于 Level-1 图片或表格形式的科学数据，利用已有技术可将其与其他密切相关的科技文献及数据进行关联，将科学数据等级从 Level-1 提升至接近 Level-2 的简单格式数据，能够为后续研究者提供更多参考信息。其次，对于 Level-2 简单格式数据，将科学数据与相关的软件、程序包建立关联，促使科学数据等级从 Level-2 提升至更高层次，使得其能够更好地支撑完整的科学分析流程。通过以上两种提高科学数据等级的可行途径，能够高效地提高科学数据可复用性，加快科学数据的共享与传播。

4.2 完善科学数据规范制度，保证科学数据基础质量

现有的科学数据管理规范制度的建设与应用尚不完善，需要对科学数据的汇交、保存、开放推广等方面进行引导与规范。具体包括：第一，科学数据管理方应明确对于提交方的要求，根据《科技计划形成的科学数据汇交技术与管理规范》 等国家标准，进一步拓展明确建立与应用高能物理科学数据标准，引导作者和出版机构补充科学数据的背景部分的介绍。针对部分重要字段缺失比例较高的问题，以及字段含义不清晰的问题，明确有价值的字段的含义，以及给出示范以提高科学数据的对应字段的准确性。这样能够使得科学数据的提交方能够按照规范标准进行科学数据的采集生成加工整理，保证所提交的科学数据的真实性、准确性、可用性、完整性、一致性、安全性等。

第二，对于科学数据管理方本身，应建立拓展高能物理科学数据质量审查制度，出具审查报告，保障汇交数据质量。在科学数据的复用方面，应充分理解数据仓储、数据研究的一出版模式对于科学数据复用的促进作用，出版模式为数据仓储和数据集及其研究的科学数据的复用性显著高于图片型表格型科学数据，可注重集约化建设高质量的数据仓储设施。以这些建议为试点，有助于明确科学数据复用的规范和要求，提高科学数据的出版质量和流通效率，促进科学数据的复用以充分挖掘科学数据的内在价值。

4.3 完善高能物理科学数据分级分类共享机制，实现高质量可持续发展

科学数据的分级分类机制是实现科学数据全生命周期安全管理的重要手段，能够保证科学数据的机密性安全要求，以及考虑不同子领域研究的需要，和不同等级的科学数据的存储管理成本。第一，对于涉及国家秘密的数据，建立数据保护指导性规范，采取禁止开放的方式。第二，对于涉及个人隐私的科学数据，可利用差分隐私等隐私保护算法进行加密技术处理。第三，对于 Level-1 等级的科普类别的科学数据，这类科学数据属于公众迫切需要的、对国家产业发展不形成竞争威胁的重要资源，应采取无偿开放的方式。第四，对于前沿类别子领域的科学数据，应优先完善提升相关服务，保证科学数据质量。对于其密切相关的科技文献、软件、程序包，开展个性化推荐工作，使得更能够支撑完整的科学分析。以这些建议为试点，有助于保证科学数据的安全需求，实现高质量可持续发展。

4.4 完善交流合作机制，推广宣传高能物理科学数据平台的使用

建议科学数据中心与相关具有科技服务优势的机构如图书馆、情报中心加强合作，加大推广平台的使用。根据本文的研究结果，高能物理科学数据的整体复用比例较低，且复用行为集中于数据仓储等完善的数据平台，而 DCI 数据库中存在着海量科学数据未被访问和复用。科技服务机构应协助科学数据中心，在保证提升科学数据等级，以及对汇交的科学数据的可访问性、可互操作性、可复用性等质量进行检查的基础之上，进一步完善科学数据的外部宣传服务。通过加大推广科学数据平台，可以提高科学数据的可发现性。进一步地，服务机构能够精准匹配相关科学领域的研究机构及科研人员，通过开展关于科学数据平台如 DCI 数据库、国家高能物理科学数据中心的培训介绍，从而促进科学数据的开放共享与复用。

5 结语

本文基于 DCI 数据库收集得到的高能物理科学数据，通过统计回归模型及多视角的描述性分析，进行属性特征和复用特征及其影响因素的实证研究。本文的研究提出了高能物理科学数据管理与复用的优化提升路径，对于提高科学数据的复用比例提供了参考。

本研究的结论也存在一些局限性和值得拓展研究的地方。第一，由于高能物理科学数据的被引频次为 0 的比例高达 86.2%，且关键词等字段缺失比例高，这使得对科学数据复用影响因素研究的结论尚不够丰富，很多特征对于被引频次的影响不显著。而对于平均被引频次较高的学科领域，如生物医学领域，可以通过探究时间段与科学数据关键词的交叉项的回归分析结果，从而探究不同时间段内各个主题对复用的影响程度，可以得到更为丰富且细粒度的影响因素的研究结论，以及更为细粒度的针对性指导建议。第二，由于不同学科的科学数据的特点往往各不相同，需要针对各学科的科学数据的特点进行特征提取，并选择所适应的统计回归模型，探究复用影响因素。

参考文献

[1] TENOPIR C,DALTON E D,ALLARD S,et al. Changes in data sharing and data reuse practices and perceptions among scientists worldwide[J]. Plos One,2015,10(8):24.

[2]杨宁,张志强. 结合计量分析和内容分析的科学数据集使用特征研究[J]. 图书情报工作,2022,66(10):122-130.

[3]孙玉伟,成颖,谢娟. 科研人员数据复用行为研究:系统综述与元综合[J]. 中国图书馆学报,2019,45(03):110-130.

[4] BORGMAN C L. The conundrum of sharing research data[J]. Journal of the American Society for Information Science and Technology,2012,63(6):1059-1078.

[5]焦红,杨波,周琪. 生物医学领域科学数据集复用特征研究[J]. 情报理论与实践,2021,44(09):90-96.

[6] PARK H,WOLFRAM D. An examination of research data sharing and re-use: Implications for data citation practice[J]. Scientometrics,2017,111(1):443-461.

[7] XIA J F,LIU Y. Usage patterns of open genomic data[J]. College & Research Libraries,2013,74(2):195-206.

[8] JOO Y K,KIM Y. Engineering researchers' data reuse behaviours: A structural equation modelling approach[J]. Electronic Library,2017,35(6):1141-1161.

[9] FANIEL I M,JACOBSEN T E. Reusing scientific data: How earthquake engineering researchers assess the reusability of colleagues' data[J]. Computer Supported Cooperative Work-the Journal of Collaborative Computing and Work Practices,2010,19(3-4):355-375.

[10] SHUTSKO A,STOCK W G. Information scientists' motivations for research data sharing and reuse[J]. Libri-International Journal of Libraries and Information Studies,2023:14.

[11]孔晔晗,张潇月,李宜展. 美国高校图书馆促进数据重用的服务实践及启示[J]. 图书与情报,2023(04):78-89.

[12]张莹,戚景琳,孙玉伟. 管理学科研人员数据复用行为特征探析[J]. 信息资源管理学报,2020,10(04):79-87.

[13]戚景琳,张莹,孙玉伟. 社会科学科研人员数据复用行为研究——以经济学为例[J]. 情报理论与实践,2020,43(09):72-78.

[14]屈宝强. 中国科学数据基础设施建设及发展对策研究[J]. 情报工程,2020,6(01):11-21.

[15] DARCH P T,KNOX E J M. Ethical perspectives on data and software sharing in the sciences: A research agenda[J]. Library & Information Science Research,2017,39(4):295-302.

[16] STVILIA B,HINNANT C C,WU S H,et al. Research project tasks, data, and perceptions of data quality in a condensed matter physics community[J]. Journal of the Association for Information Science and Technology,2015,66(2):246-263.

[17]孙玉伟. 基于“方差”和“过程”理论的科研人员数据复用行为研究逻辑框架[J]. 图书馆学研究,2020(18):70-79.

[18]黄如花,林焱. 法国政府数据开放共享的政策法规保障及对我国的启示[J]. 图书馆,2017(03):1-6.

[19]孙浩,陈美. 荷兰政府开放数据的政策法规保障及启示[J]. 情报杂志,2021,40(02):161-168.

[20]陈昕,郑晓欢,潘博雅,et al. 中国科学院科学数据中心体系建设实践及展望[J]. 中国科学数据(中英文网络版),2023,8(01):146-164.

[21] CAO Y,JONES C,CUEVAS-VICENTT ́N V,et al. Dataone: A data federation with provenance support[C].6th International Provenance and Annotation Workshop (IPAW),2016: 230-234.

[22] ABELLA A,ORTIZ-DE-URBINA-CRIADO M,DE-PABLOS-HEREDERO C. Open data quality metrics: Barcelona open data portal case[J]. Profesional De La Informacion,2018,27(2):375-382.

[23] AKERS K G,READ K B,AMOS L,et al. Announcing the <i>journal of the medical library association</i>'s data sharing policy[J]. Journal of the Medical Library Association,2019,107(4):468-471.

[24] ROA-MART ́NEZ S M,VIDOTTI S A B,SANTANA R C. Proposed structure of a data paper structure as scientific publication[J]. Revista Espanola De Documentacion Cientifica,2017,40(1).

[25] TENOPIR C,SANDUSKY R J,ALLARD S,et al. Research data management services in academic research libraries and perceptions of librarians[J]. Library & Information Science Research,2014,36(2):84-90.

[26] PIWOWAR H A,VISION T J. Data reuse and the open data citation advantage[J]. Peerj,2013,1.

[27]陈铭. 知识生产视域下的科学数据出版实践——兼论学术出版商的角色定位与功能分析[J]. 科技与出版,2022(12):102-109.

[28]黄欣卓,米加宁,章昌平,等. 科学数据复用研究的演化、知识体系与方法工具——兼论第四科研范式的影响[J]. 科研管理,2022,43(08):100-108.

[29] CURTY R G,CROWSTON K,SPECHT A,et al. Attitudes and norms affecting scientists' data reuse[J]. Plos One,2017,12(12).

[30]商雯雯. 基于元分析的科学数据重用意愿影响因素研究[D].2022.
[31]文静,何琳,韩正彪. 科研人员科学数据重用意愿的影响因素研究[J]. 图书情报知识,2019(01):11-20.
[32]商雯雯,支凤稳,孟佳琪. 科学数据重用意愿影响因素的元分析[J]. 晋图学刊,2023(05):1-17+28.
[33]王雪,杨波. 科学数据重复使用的学科差异性研究[J]. 情报杂志,2021,40(07):122-126+156.
[34]盛小平,袁圆. 科学数据开放共享中的数据权利治理研究[J]. 中国图书馆学报,2021,47(05):80-96.
[35]罗鹏程,崔海媛,赵静茹. 基于 datacite 的科学数据现状特征研究[J]. 图书情报知识,2019(03):101-112+80.
[36]石京燕,黄秋兰,汪璐,et al. 国家高能物理科学数据中心分布式数据处理平台[J]. 数据与计算发展前沿,2022,4(01):97-112.
[37]曾珊,陈刚,齐法制,等. 高能物理科学数据服务与应用[J]. 中国科技资源导刊,2022,54(01):31-39.
[38]杨宁,文奕,张鑫,等. 高能物理科学数据与科技文献关联研究[J]. 图书馆学研究,2019(01):47-52.
[39]齐法制,陈刚,程耀东. 建立权责明晰且能力健全的科学数据开放共享机制——以高能物理领域为例[J]. 中国科学基金,2019,33(03):229-236.
[40] DALLMEIER-TIESSEN S,MELE S. Integrating data in the scholarly record: Community-driven digital libraries in high-energy physics[J]. Zeitschrift Fur Bibliothekswesen Und Bibliographie,2014,61(4-5):220-223.
[41] SHIERS J. Dphep: From study group to collaboration[J]. 20th International Conference on Computing in High Energy and Nuclear Physics (Chep2013), Parts 1-6,2014,513.
[42] ZEILEIS A,KLEIBER C,JACKMAN S. Regression models for count data in r[J]. Journal of Statistical Software,2008,27(8):1-25.
[43]杨宁,张志强. 科学数据集知识扩散特征探析——以基因表达数据集为例[J]. 图书情报工作,2022,66(12):82-91.

作者简介：胡威，男，1997 年生，博士，特别研究助理，研究方向：网络结构数据、空间计量模型、科学计量学。
杨宁，男，1982 年生，博士，副研究馆员，研究方向：学科信息学，科学数据管理与评价。通信作者：杨宁(yangn@clas.ac.cn)

作者贡献声明: 胡威，提出研究思路，数据采集，数据分析，论文组织与撰写。 杨宁，提出研究选题和思路，数据处理，论文修改。